TR02: State dependent oracle masks for improved dynamical features

J. F. Gemmeke and B. Cranen

Dept. of Linguistics, Radboud University, Nijmegen, The Netherlands
{J.Gemmeke, B.Cranen}@let.ru.nl

Abstract

Using the AURORA-2 digit recognition task, we show that recognition accuracies obtained with classical, SNR based oracle masks can be substantially improved by using a state-dependent mask estimation technique.
Index Terms: Noise Robust ASR, Missing Data Techniques

1. Introduction

In Missing Data Techniques (MDT) for noise robust automatic speech recognition (ASR), it is often implicitly assumed that using an SNR based oracle mask [1] guarantees maximum recognition accuracy. Generally speaking, however, this is not necessarily true.

In previous work [1], the authors showed that the portions of an oracle mask which are important for recognition accuracy are speech sound dependent. In this paper we exploit this finding by a state dependent treatment of reliable features. Using a different mask estimator for every state in an HMM model and selecting the mask estimator for each time frame based on an externally provided state transcription, we generate a new type of oracle mask. In this paper we compare recognition accuracies obtained with state-dependent oracle masks and classical oracle masks on the AURORA-2 digit recognition task.

2. Experiments and results

Experiments on test set A of the AURORA-2 digit corpus were carried out using a MATLAB HMM-based MDT recognizer in which the masks for delta coefficients were computed as the delta's of the static masks (cf. [2] for implementation and model details). Our state-dependent mask estimator was based on binary SVM classifiers using LIBSVM.

We trained separate SVM-models for all $S = 179$ HMM states and all $K = 23$ Mel frequency bands, resulting in $S \times K = 4117$ models. The frame-based SVM features we used consisted of 'Subband Energy to Subband Noise Floor Ratio' and 'Flatness' as in [3], the harmonic and random components of the noisy speech signal [4] and the noisy speech acoustic vectors. Reliability labels used in training were obtained from the (classical) oracle mask. Every state-specific SVM mask estimator was trained on the frames from the multi-condition train set, which were assigned to the same state by a forced alignment of the corpus utterances with the reference transcription. All 4117 models were trained with the same SVM-feature vector.

Table 1 shows the recognition accuracies for the classical and the state-dependent oracle mask, as well as the accuracy gain.

[1] These oracle masks are computed by comparing spectro-temporal representations of the underlying speech and noise signals. Features dominated by speech energy are dubbed reliable; features dominated by noise energy unreliable.

Table 1: AURORA-2 digit recognition accuracy (%).

	SNR						
method	clean	20	15	10	5	0	-5
classical	99.7	99.2	99.3	98.4	96.3	88.3	58.6
state dep.	99.7	99.5	99.5	99.2	97.9	92.1	67.3
Δ acc.	0.0	0.3	0.2	0.8	1.5	3.8	8.7

Clearly, the state dependent method performs consistently better than the classical oracle mask, with larger gains in more adverse conditions.

3. Discussion and Conclusions

The classical, SNR based oracle mask only describes which static coefficients are reliable. Since treatment of dynamic features is missing data decoder specific the classical oracle mask is not necessarily the 'ideal' mask. Detailed analysis revealed that state dependent masks contain fewer isolated reliable elements than classical ones. In our setup (but also in those of others, e.g., [5]) coarser granularity is beneficial for recognition performance, because isolated reliable mask coefficients can result in delta mask coefficients that are mistakenly labeled reliable.

Our findings might also be useful for speech decoding without a priori information about the state sequence. Oracle recognition accuracy would theoretically come within reach if, for each frame, one can afford to evaluate as many mask vectors as there are states (i.e. 179 in case of AURORA-2).

This is a significant reduction of complexity as compared to the $2^{23} = 8.388.608$ theoretically possible masks and without the loss of accuracy reported in [5]. For small vocabulary tasks, today's computing power might even make a brute force approach feasible.

Our future research, however, will focus on a further reduction of computational complexity by exploiting state transition constraints.

4. Acknowledgements

The research of Jort Gemmeke was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program.

5. References

[1] J. Gemmeke, B. Cranen, and L. ten Bosch, "On the relation between statistical properties of spectrographic masks and recognition accuracy," in *SPPRA - 2008*, 2008, pp. 200–206.

[2] H. Van hamme, "Handling time-derivative features in a missing data framework for robust automatic speech recognition," in *Proceedings of IEEE ICASSP*, 2006.

[3] M. Seltzer, B. Raj, and R. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, 2004.

[4] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proceedings of IEEE ICASSP*, vol. 1, 2004, pp. 213–216.

[5] S. Demange, C. Cerisara, and J. Haton, "Missing data mask models with global frequency and temporal constraints," in *Interspeech-2006*, 2006.